|  |
| --- |
| **HR Analytics - 누구를 뽑을 것인가?**  **1. 김소정(\*\*\*\*\*\*\*\*)**  **2. do\_somting@naver.com**  **3. 김소정(\*\*\*\*\*\*\*\*)/이주영(\*\*\*\*\*\*\*\*)/김민지(\*\*\*\*\*\*\*\*)** |

'나는 IBM 회사의 대표이다.

현재 하반기 신입사원 채용 최종 컨펌을 앞두고 고민하고 있다.

사실 어느 부서에 속한 직원이라도 직원들의 이직은 회사의 입장에서 큰 손실이다.

비슷해 보이는 스펙의 지원자 중에 어떤 사람이 회사를 떠나지 않을까에 대한 기준을 마련할 수 있을까?'

한 예로 프린터 제조업체 제록스는 콜센터 직원을 선발할 때 관련 경험을 기준으로 선발을 하였다. 그러나 콜센터의 특성상 직원의 조기 퇴사율이 높았고, 1인당 5000달러에 해당하는 훈련교육비 손실이 발생하였다. 그렇다면 조기 퇴직자를 사전에 예측하면 어떨까? 직원들을 대상으로 심리테스트를 진행한 결과 조기 퇴사자의 특징은 회사에서 멀리 거주하며, 확실한 교통 수단이 없고 외톨이거나 5개 이상의 소셜 네트워크 관련하고 궁금한 것이 너무 많고 공감을 너무 잘하거나 창의력이 부족한 것으로 나타났다. 그래서 제록스는 이를 근거로 고용 기준 변경을 결정하였고 그 결과 6개월 후 퇴사율은 20% 감소하였다.

좋은 인재를 찾아서 채용하는 것은 중요하다. 하지만 이들의 이직을 막는 것 또한 중요하기에 우리는 직원의 이직 가능성을 예측하는 방법에 대해서 분석해보고자 한다. 그래서 우리는 어떠한 변수가 직원들의 이직률에 영향을 끼치는지를 알아보고 이를 예측할 수 있는 모델을 세워보고 결과를 해석할 것이다.

분석에는 Kaggle의 ‘IBM HR Analytics Employee Attrition & Performance’의 dataset을 이용하였다.



- 데이터 출처 : <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

- IBM사의 2940명의 직원들에 대한 35가지의 다양한 변수 데이터

**어떠한 변수들이 있는가?**

35가지의 변수들은 다음과 같다.

|  |  |
| --- | --- |
| 1. Attition : Yes/No 직원들의 이직 여부  2. BusinessTravel  : Non-Travel/Travel\_Rarely/Travel\_Frquntly 출장빈도  3. DailyRate  4. Departmet : Sales/Research & Development/Human Resources 소속  5. DistanceFromHome  6.Education : 교육수준,1-Below College,2-College,3-Bachelor,4-Master,5-Doctor  7.EducationField : Human Resources/Life Sciences/Marketing/Meical/Other/Technicl Degree  8.~~EmployeeCount~~  9.~~EmployeeNumber~~ : id  10.EnvironmetSatisfaction  11.Gender  12.HourlyRate  13.JobInvolvemet : 직업 관여도(1~4)  14.JobLevel : 직위(1~5)  15.JobRole : 직함  16.JobSatisfaction | 17.MaritalStatus : Single/Married/Divorced  18.MonthlyIncome  19.MonthlyRate  20.NumCompaniesWorked  21.~~Over18~~ : all yes  22.OverTime : Yes/No 초과근무여부  23.PercentSalaryHike : 연봉상승률  24.PerformanceRating : 성과평가  25.RelationshipSatisfaction  26.~~StandardHours~~ : all 80  27.StockOptionLevel  28.TotalWorkingYears  29.TrainingTimesLastYear  30.WorkLifeBalance  31.YearsAtCompany  32.YearsInCurrentRole  33.YearsSinceLastPromotion  34.YearsWithCurrManager  35.Age |

이 중 over18, standarthours, employeecount, employeeNumber는 불필요한 정보로 판단되어 본 분석에서는 해당 변수를 제외하고 31가지의 변수를 사용하였다.

**데이터탐색**

**1. 개괄탐색**

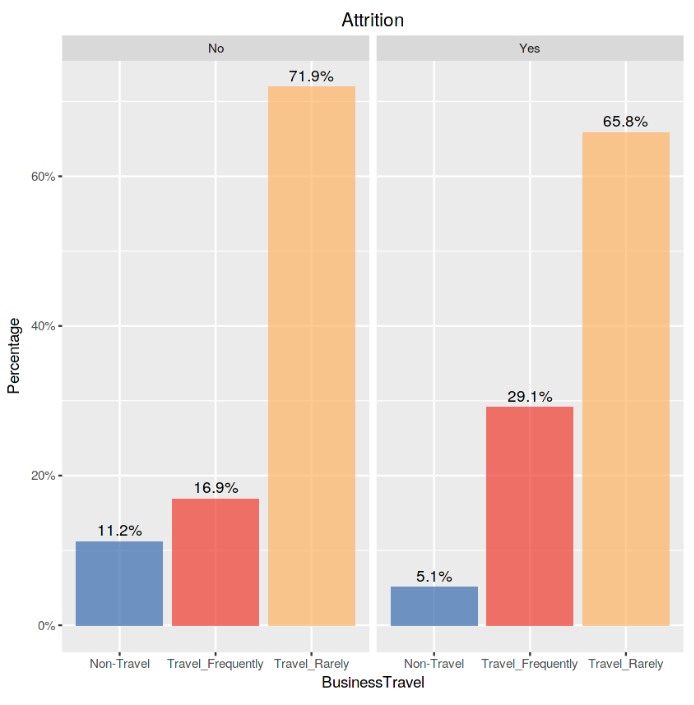
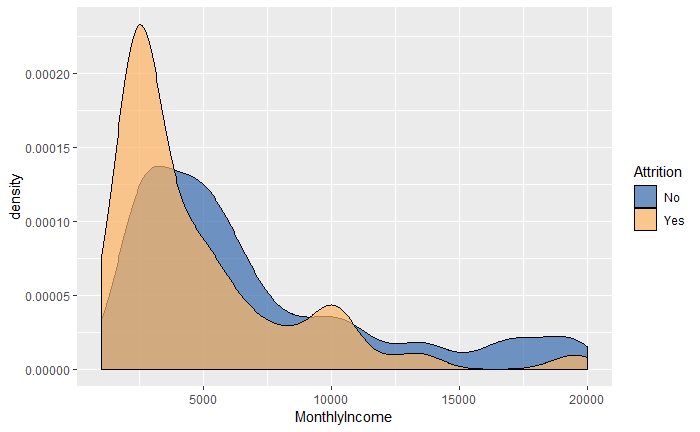
**인구 통계학적 변수들을 가지고 탐색해 보았다. 이직 여부로 나누어 살펴보겠다.**

**1-1 나이와 결혼여부**

|  |  |
| --- | --- |
|  |  |

20대 중반에서 40대 중반의 근로자가 대부분이다. 그 중 퇴사한 사람 중 20에서 30대 사람이 훨씬 많았다. 젊은 사람이 퇴사를 많이 하는 것으로 보인다. 20대 중반에서 40대 중반의 근로자가 대부분이다. 그 중 퇴사한 사람 중 20에서 30대 사람이 훨씬 많았다. 젊은 사람이 퇴사를 많이 하는 것으로 보인다. 퇴사한 사람 중 싱글인 사람이 훨씬 높은 비율을 보였다. 회사에 근속한 사람은 결혼한 경우가 가장 많았다.

**1-2 월급과 여행 여부**

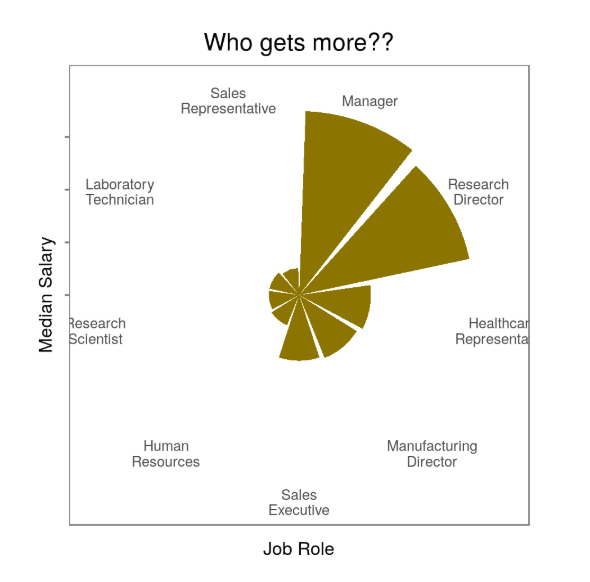
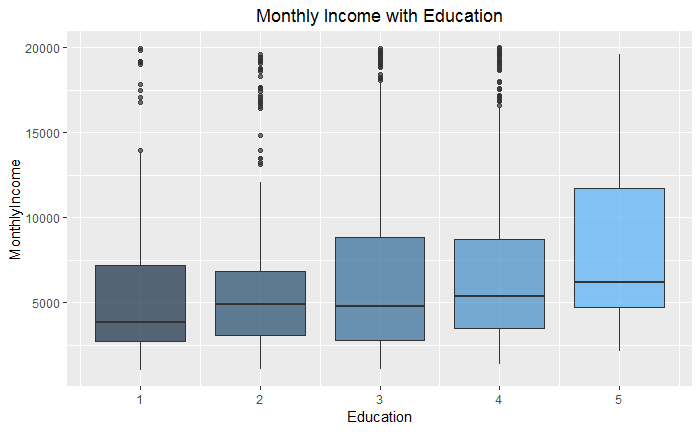


월급의 경우 대체적으로 5000달러 전후가 가장 많았으나 퇴사하는 사람의 월급은 1500달러에 집중되어있다. 퇴사하는 사람의 월급이 훨씬 적은 경향이 있다. 퇴사한 사람의 경우, 출장의 경우 퇴사한 사람이 더 출장을 가는 경향이 있다.

**2. 어떤 사람에게 돈을 많이 줘야 할까?**

**직원의 임금을 얼마나 줘야 할 것인지도 기업의 큰 이슈이다. 과하게 주면 기업의 손해이고 직원의 능력보다 적게 주면 우리의 손실일 것이다. 월급인 Monthly Income 변수와 다른 변수와의 관계를 살펴보겠다.**

**2-1 학력과 직군**



학력이 높을수록 월급이 높은 경향을 보인다. College의 경우 분포가 많아서 높게 나온것으로 보인다. 직군 별로 보았을 때, Manager 직군과 Research Director직군의 월급이 타 직군의 약 7배 정도 높았다.

**3. 이직에 영향을 주는 요인은 무엇이 있을까?**

**미국 매체의 조사에 따르면 이직 여부는 직무 적합성, 회사의 문화, 보상의 적합성 등 3가지 요인으로 나뉜다고 한다. 이에 따라 변수들을 정성적으로 분류해보고 어떤 요인이 이직에 영향을 주는지 탐색해 보았다. 변수는 다음과 같이 분류하였다. 인구통계학적 변수는 앞에서 다루었으므로 넘어가겠다.**

1. 인구 통계학적, 개괄 변수 - age, business travel, Education,Gender ,MaritalStatus

2. 직무 적합성 - Department, EducationField, JobInvolvemen, JobRole, JobSatisfaction, PerformanceRating

3. 직업 환경 적합성

3-1 회사 문화, 동료 만족도 - EnvironmentSatisfaction, RelationshipSatisfaction

3-2 워라벨 만족도 - DistanceFromHome, OverTime, WorkLifeBalance

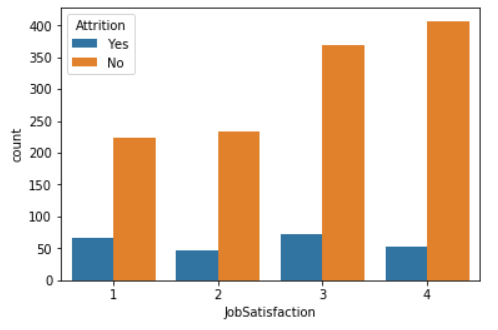
4. 대우 적합성

4-1 임금에 대한 만족도 - DailyRate ,HourlyRate, MonthlyRate, MonthlyIncome, StockOptionLevel

4-2 진급에 대한 만족도 - PercentSalaryHike, JobLevel

5 . 경력 관련 만족도 - NumCompaniesWorked, Years AtCompany, TotalWorkingYears, Years SinceLastPromotion (years와 관련된 변수이므로 따로 분류하였다.)

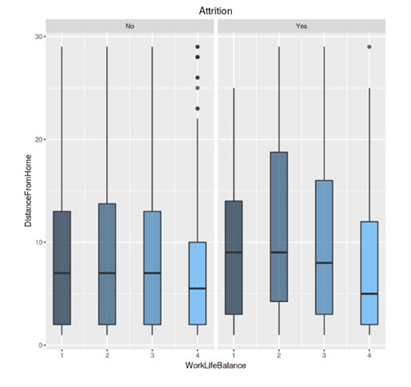
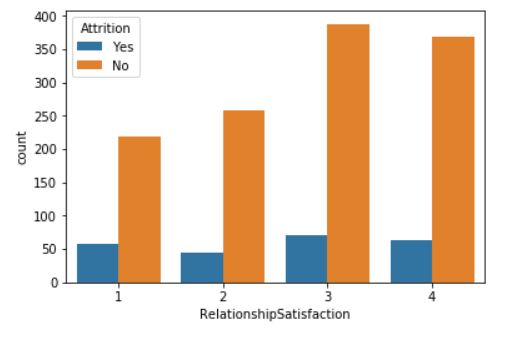
**3.1. 직무적합성**



일에 만족하지 않는 사람이 퇴사할 비율이 높게 나타난다.

**3.2. 직업환경 적합성**

**3.2.1 - 3.2.2 회사 문화 동료 만족도, 워라벨 만족도**

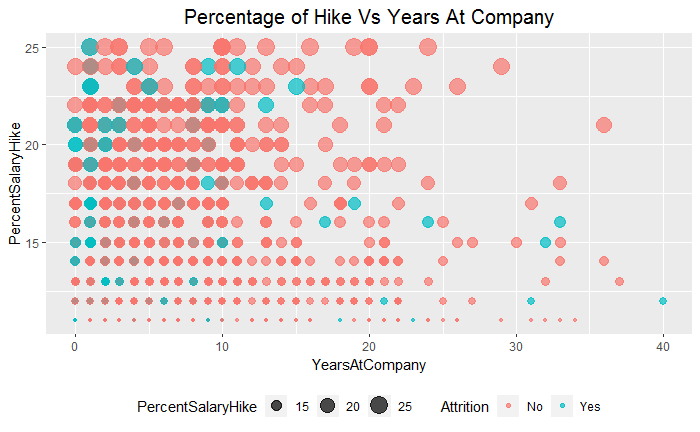
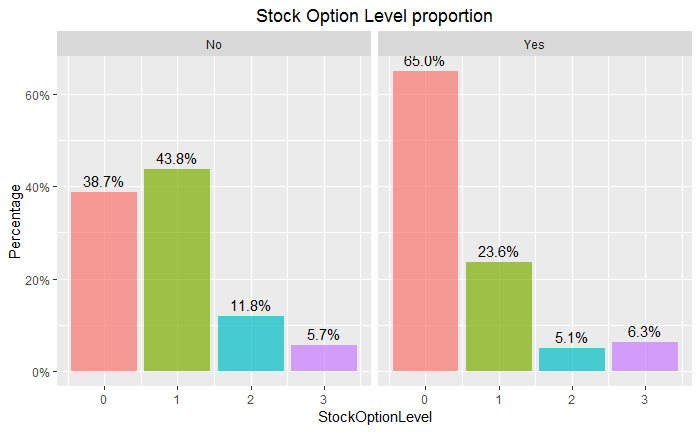


동료와의 관계가 좋지 않다고 생각하는 사람이 퇴사하는 비율이 더 높다.

근속하는 사람의 경우 워라벨과 상관없이 회사에서 집까지 거리 평균에 큰 차이가 없는 것으로 보아 워라벨에 집까지의 거리가 영향을 주지 않는 것으로 보인다. 그러나 퇴사한 사람의 경우 워라벨이 낮을수록 집까지의 거리가 먼 추세를 보인다.

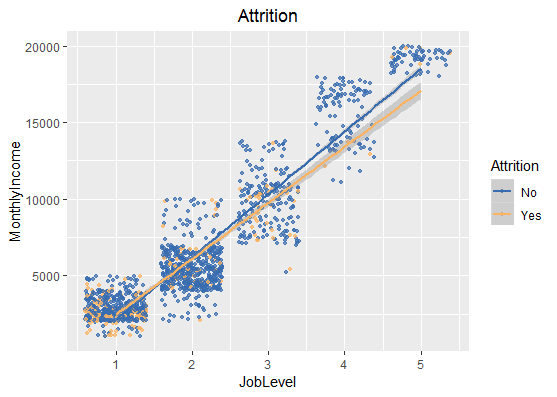
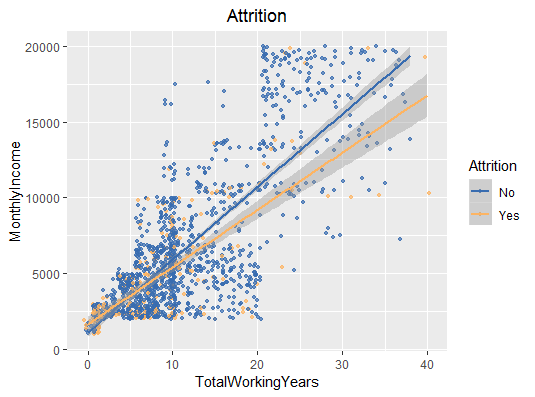
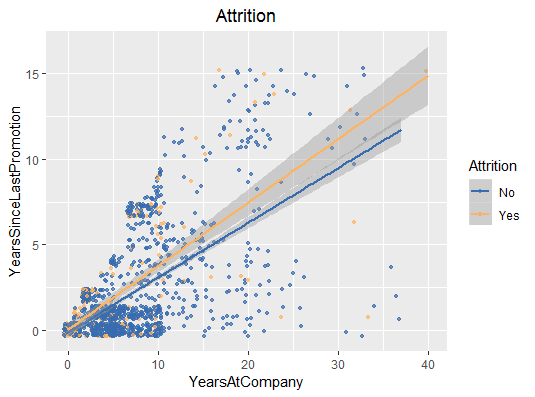
**3.3 대우적합성**

**3.3.1-3.3.2 임금에 대한 만족도와 진급에 대한 만족도**



회사 주식 소유 여부는 소속감과 직결되기도 한다. 회사 주식을 소유하지 못한 사람의 경우 퇴사하는 비율이 크게 나타났다. 회사 근속 연수와 임금 상승률을 퇴사 여부로 보았을 때, 퇴사 하는 사람은 근속 연수가 짧을 수록 연봉상승률이 높게 나타났다. 이는 호봉제인 우리나라와 달리, 직장을 옮길 때 연봉을 올리는 미국의 특성이 드러난 것 같다. 퇴사하지 않은 사람보다 이러한 경향이 더 두드러 지는 것을 보아 근속에 비해 적절한 보상을 받지 못한다고 생각할 수 있다.

**3.4 경력 관련 만족도**



1사분면에 해당하는 표를 보면 이직하는 사람은 해당 기업 근속 연수에 비해 진급이 더디다. 2사분면에 해당하는 총 근무 연수와 월급은 양의 상관관계를 지닌다. 이직하는 사람은 근무 연수보다 적은 월급을 받는 경향이 있다. 3사분면에 해당하는 표의 경우, 직급에 따라 월급이 높으며 퇴사하는 사람의 월급이 더 낮은 경향을 보인다.

**예상했듯 직무에 잘 맞지 않고, 동료와 워라벨에 불만이 있는 사람, 임금과 진급에 대한 만족이 낮은 사람이 퇴사할 가능성이 컸다.**

**변수 사이의 연관성을 분석할 수 있을까?**

앞서 우리는 31개의다양한 변수들의 관계를 시각적으로 확인 해 보았다. 이 변수들의 차원을 축소해서 다룰 수는 없을까? 앞서 상관관계를 살펴본 수치형 변수 15개에 대해 주성분분석(PCA)과 요인분석(Factor Analysis)을 사용하여 차원 축소의 가능성에 대해 생각해 보았다.

**수치형 변수들의 차원 축소해보기 – 주성분분석(PCA)**

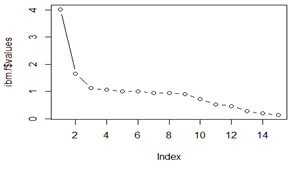
주성분 분석을 이용해 기존 변수들의 분산, 변이를 최대한 설명하는 서로 독립인 주성분을 찾아내자. 주성분 분석의 결과로 얻는 주성분들은 기존 변수의 선형 결합으로 표현된다.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 |
| 변동  설명 | 0.268 | 0.11 | 0.075 | 0.071 | 0.068 | 0.067 | 0.064 | 0.063 | 0.06 | 0.048 | 0.035 | 0.031 | 0.019 | 0.013 | 0.009 |
| 누적  비율 | 0.268 | 0.378 | 0.453 | 0.523 | 0.591 | 0.657 | 0.721 | 0.784 | 0.844 | 0.892 | 0.928 | 0.959 | 0.978 | 0.991 | 1 |

위 표는 15개의 수치형 변수들을 주성분분석한 결과로, 주성분들이 설명하는 변동의 누적을 보여준다. 총 변동의 95%를 설명하기 위해서는 12개의 변수를, 90%를 설명하기 위해서는 11개의 변수를 사용하여 차원을 축소할 수 있다. 누적 분포를 보면 전체 변동이 여러 변수에 퍼져 있어 차원 축소가 효율적으로 이루어지지 않는다는 것을 확인하였다.

**수치형 변수들의 잠재된 요인 찾아보기 – 요인분석(FactorAnalysis)**

차원 축소의 다른 방법으로는 요인분석(Factor Analysis)이 있다. 요인분석은 여러 개의 변수가 있을 때, 이들을 설명하는 잠재된 요인을 찾아내고자 하는 것이다.



위 그래프는 측정 변수들 간의 상관관계에 따라 상관계수행렬의 고유근을 나타낸 것이다. 요인의 수는 해당 값이 1 이상인 경우 채택하여 6개로 결정하였다. 다음은 6개의 요인을 이용한 요인분석의 결과이다.

Factor1 Factor2 Factor3 Factor4 Factor5 Factor6

SS loadings 1.748 1.472 1.031 1.010 0.998 0.207

Proportion Var 0.117 0.098 0.069 0.067 0.067 0.014

Cumulative Var 0.117 0.215 0.283 0.351 0.417 0.431

Test of the hypothesis that 6 factors are sufficient.

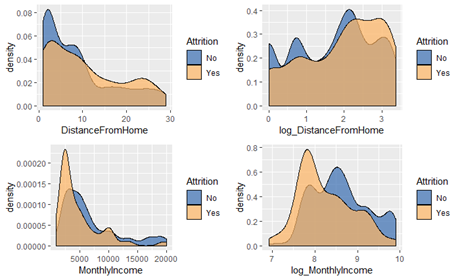
The chi square statistic is 17.66 on 30 degrees of freedom. The p-value is 0.964

SS loadings는 기존 변수와의 상관관계 값인 로딩 값의 제곱 합에 해당하는 것으로 Factor 6의 경우 그 값이 작아 변수를 잘 설명하지 못하는 편이다. 6개의 요인은 원래 분산의 43.1% 가량을 설명한다. 그러나 가설 검정 결과는 6개의 요인이 모델을 설명하기에 충분하다는 것을 지지한다.

**이직 여부를 설명하는 모형을 어떻게 세울까?**

**Logistic Regression**

앞서 변수들의 차원을 축소하려는 시도를 했지만 위 분석은 연속형 변수에 대해서만 진행했고, 차원의 축소가 효과적으로 이루어지지는 않았다. 우리는 새로운 사원의 이직 가능성을 살펴보고 싶기 때문에 로지스틱 회귀모형을 사용하여 이직율을 예측하는 새로운 모형을 세우고자 한다. 범주형 변수는 dummy variable로 처리했고, 0 부근에 치우쳐져 있는 DistanceFromHome과 MonthlyIncome의 경우 정규화를 위해 로그를 취했다.



모든 변수를 사용할 경우 과적합의 오류가 생길 가능성이 있으므로 AIC를 가장 작게 하는 방향으로의 단계적 변수 선택을 진행해 최적의 모델을 생성하였다. 여기서 AIC를 최소화 하는 것은 우도(Maximum Likelihood)를 최대화하면서 변수의 개수를 줄이는 방향을 취했다. 이 결과 다음의 19개 변수를 지니는 모델이 생성되었다.

Attrition ~ Age + BusinessTravel + Department +

EducationField + EnvironmentSatisfaction +Gender + JobInvolvement +

JobSatisfaction + MaritalStatus +NumCompaniesWorked + OverTime +

RelationshipSatisfaction +TrainingTimesLastYear + WorkLifeBalance +

YearsInCurrentRole + YearsSinceLastPromotion +YearsWithCurrManager+

log\_DistanceFromHome + log\_MonthlyIncome

Null deviance: 974.94 on 1102 degrees of freedom

Residual deviance:648.02 on 1076 degrees of freedom

AIC: 702.02

19개의 변수 또한 과적합일 가능성이 있어, AIC가 가장 적게 증가하는 방향으로 추가적으로 변수를 제거하여 Reduced Model을 생성했다. 결과적으로 선택된 10개의 변수와 최종 식은 다음과 같다.

Model : Attrition ~ BusinessTravel + Department +EnvironmentSatisfaction +

JobInvolvement + MaritalStatus + OverTime + YearsSinceLastPromotion +

YearsWithCurrManager + log\_DistanceFromHome + log\_MonthlyIncome

Null deviance: 974.94 on 1102 degrees of freedom

Residual deviance:716.66 on 1089 degrees of freedom

AIC: 744.66

Logit(이직할 확률)=9.56+1.52\*BusinessTravel\_Frequently+0.91\*BusinessTravel\_Rarely-0.60\*Department\_RnD

+0.26\*Department\_Sales-0.49\*EnvironmentSatisfaction-0.55\*JobInvolvement+0.32\*MaritalStatus\_Married+1.23\*MaritalStatus\_Single+1.76\*OverTime\_Yes+0.16\*YearsSinceLastPromotion-0.16\*YearsWithCurrManager+0.33\*log\_DistanceFromHome

-1.29\*log\_MonthlyIncome

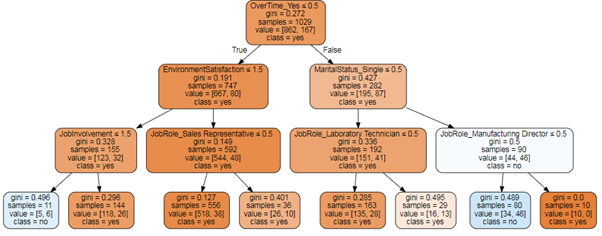
변수의 개수를 각각 31개, 19개, 10개 사용한 모델에 따른 테스트 셋 일치 정도는 다음과 같았다.

|  |  |  |  |
| --- | --- | --- | --- |
| 변수 개수 | 31개  (full model) | 19개 | 10개  (reduced model) |
| 정확도 | 0.886 | 0.88 | 0.858 |

모형의 결과 full model이 테스트 집합을 가장 잘 설명하였으나 과적합을 방지하기 위해 세운 19개, 10개의 모형도 테스트 집합을 잘 설명하는 것을 확인했다.

**Decision Tree**

명목형 변수를 dummy 변수화 하여 Decision Tree 모델에 적용해 보았다. Entropy 지수를사용한 경우보다 geni 계수를 적용한 모델이 정확도가 0.2 정도 더 높아( 84.3537), 이 모델을 선택하였다.



퇴사하는 사람들은 추가 근무를 할 때, 환경 만족도가 낮을 때, 직업 관여도가 낮을 때 영업 직무에 속할 때 경향이 두드러진다. 회사에 근속하는 사람들은 추가 근무를 하지 않는 경향이 있으며 싱글이고, 연구개발직군, 생산관리 직군인 경향이 있다.

**누가 뽑힐 것인가?**

우리는 위의 모형 중에서 정확도가 가장 높은 Logistic Regression을 최종 모형으로 선택하였다.

Logit(이직할 확률)=9.56+1.52\*BusinessTravel\_Frequently+0.91\*BusinessTravel\_Rarely-0.60\*Department\_RnD

+0.26\*Department\_Sales-0.49\*EnvironmentSatisfaction-0.55\*JobInvolvement+0.32\*MaritalStatus\_Married+1.23\*MaritalStatus\_Single+1.76\*OverTime\_Yes+0.16\*YearsSinceLastPromotion-0.16\*YearsWithCurrManager+0.33\*log\_DistanceFromHome

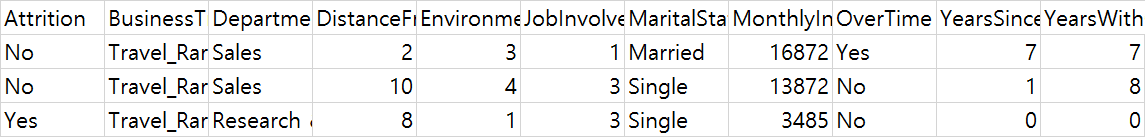
-1.29\*log\_MonthlyIncome

최종 모형에 따르면 이직한 사람의 특징은 다음과 같다.

출장을 종종 가는 경우, 부서가 R&D가 아닌 경우, 부서가 Sales인 경우, 환경에 만족하지 않고 직업관여도가 낮은 경우, 결혼을 한 경우 혹은 싱글인 경우, 또한 초과근무를 한 경우, 진급한 연수가 길수록, 사수와 일한 연수가 짧을수록, 집에서 직장까지의 거리가 멀수록, 월급이 적을수록 이직할 확률이 높다.

마지막으로 우리는 실제상황에 적용해보려고 한다.

실제 3명의 지원자가 신입사원 채용 최종면접 단계에 올라왔다. 지원자들의 특성은 다음과 같다.



위의 모형을 이용한 결과를 토대로 우리는 이직할 확률이 낮은 첫번째, 두번째 지원자를 채용할 것이다.

이처럼 우리는 직원의 이직가능성을 판단하여 채용 프로세스에 도입할 수 있으며, 이로 인한 이직률 감소를 통해 상당한 비용절감을 얻을 수 있을 것이다.